(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

J. [US/US]; 66 Maybury Hill Road, Princeton, NJ 08540 (US). OUYANG, Ming [—/US]; 9 Stoecker Road, Holmdel, NJ 07733 (US). LIOY, Paul [US/US]; 111 Holly Street, Canford, NJ 07016 (US). GEORGOPOULOS, Panos [GR/US]; 70 Spring Wood Court, Princeton, NJ 08540 (US).

(54) Title: SYSTEMS AND METHODS FOR MICROARRAY DATA ANALYSIS

```
EM_ESTIMATE (μ₁, ..., μ_K, Σ₁, ..., Σ_K, τ₁, ..., τ_K, A')
{       FOR EACH ROW R OF A' WITH MISSING VALUES
{       FOR i = 1, ..., K
{       USE EM AND N(μ_i, Σ_i) TO ESTIMATE THE
        MISSING VALUES IN R.
R_i ← R WITH MISSING VALUES REPLACED BY ESTIMATES.
        }
    }
R' ← WEIGHTEDAVERAGE(R₁, ..., R_K).
REPLACE R IN A' BY R'.
RETURN A'.
}


K_ESTIMATE(K, A)
{       /* FIRST PART: INITIALIZATION */
B ← ROWS OF A WITHOUT MISSING VALUES.
μ₁, ..., μ_K, Σ₁, ..., Σ_K, τ₁, ..., τ_K ←
    GAUSSIAN MIXTURE CLUSTERING OF B.
A' ← EM_ESTIMATE (μ₁, ..., μ_K, Σ₁, ..., Σ_K, τ₁, ..., τ_K, A).
/* SECOND PART: ITERATION */
REPEAT
{       μ₁, ..., μ_K, Σ₁, ..., Σ_K, τ₁, ..., τ_K ←
    GAUSSIAN MIXTURE CLUSTERING OF A'.
A' ← EM_ESTIMATE (μ₁, ..., μ_K, Σ₁, ..., Σ_K, τ₁, ..., τ_K, A').
} UNTIL CONVERGENCE
}


GMCimpute(S, A)
{       FOR K = 1, 2, ..., S
{       A_K ← K_ESTIMATE(K, A).
}
RETURN (A₁ + A₂ + ... + A_S) / S.
}
```

(57) Abstract: Clustering is routinely applied in the exploratory analysis of microarray data. Missing entries arise from blemishes on the microarrays. The present invention provides a new method, and computer program and/or computer product thereof to impute missing values. The method involves the steps of clustering microarray data by partitioning the data into a select number of clusters, wherein each data point is iteratively moved from one cluster to another, until two consecutive iterations have resulted in the same partition pattern; obtaining a select number of estimates of the data in the clusters by probabilistic interference; and averaging the select number of estimates to obtain missing values in the microarray data. The method is superior to other imputation models as measured by root mean squared errors.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.